SPECIFICATION AND STANDARDIZATION OF DATA ELEMENTS, RULES, AND GUIDELINES FOR THE FORMULATION OF DATA DEFINITIONS

1. PURPOSE. The policy establishes guidelines for the formulation of data definitions for the Environmental Protection Agency (EPA) Data Standards Program. The purpose of this program is to facilitate the sharing of data, to set forth Agency principles on data standards, and to assign organizational responsibilities for implementing and administering common data standards. This policy, in accordance with the International Standards, provides rules and guidelines for the construction of well-formed data element and data element concept definitions to ensure consistency and quality of a data element and data element concept.

2. SCOPE AND APPLICABILITY. This policy applies to all EPA organizations and their employees. It also applies to the facilities and personnel of agencies of EPA who design, develop, operate, or maintain Agency information and information systems. This policy applies to automated and manual systems developed for programs or administrative purposes. The requirements of this policy apply to existing data elements and data element concepts, as well as new data elements and data element concepts. Although these definitional rules and guidelines pertain to data elements, they can be applied in formulating definitions for other types of data constructs such as entity types, relationships, attributes, object types, segments, data composites, codes, and datasets.

3. BACKGROUND.

    a. Integration of information and databases is difficult because program offices use disparate formats and names for similar data elements and data element concepts.

    b. To support effectively the use of common definitions of environmental data with State programs, EPA must have common definitions for data elements and data element concepts and an intra-agency capability to share data.

    c. Precise and unambiguous data element and data element concept definitions are one of the most critical aspects of ensuring data shareability. One of the primary vehicles for carrying the data's meaning is the data element and data element concept definition. Therefore, it is mandatory that every data element and data element concept have a well-formed definition: one that is clearly understood by every user and by recipients of shared data. Poorly formulated data element and data element concept definitions foster misunderstandings and ambiguities and often inhibit successful communication.

4. AUTHORITIES. The following guidelines contain provisions that, through reference in this text, constitute provisions of this Part of the International Standard. All guidelines are subject to revision, and parties to agreements based on this Part of the

International Standard are encouraged to apply the most recent editions of the guideline indicated below. Members of the International Electrotechnical Commission (IEC) and the International Organization for Standardization (ISO) maintain registers adhering to currently valid International Standards.

References:

a. ISO/IEC 11179-3:1996, *Information technology - Specification and Standardization of Data Elements - Part 3: Basic Attributes of Data Elements*.

b. ANSI X3.172-1990, *American National Standard Dictionary for Information Systems (ANDIS)*.

c. ISO/IEC 11179-4:1996, *Information technology - Specification and Standardization of Data Elements - Part 4: Rules and Guidelines for the Formulation of Data Definitions*.

5. POLICY. It is EPA policy to create and maintain consistency between data elements and data element concepts that have more than one application within the Agency. This consistency will permit data sharing necessary to achieve environmental results. The consistency will be facilitated by precise and unambiguous data element and data element concept definitions.

6. RESPONSIBILITIES.

a. The Office of Information Resources Management (OIRM) shall:

(1) Provide effective leadership in developing, promulgating, and enforcing the Agency data element and data element concept definition guidelines.

(2) Publish data element and data element concept definitions in the Environmental Data Registry.

b. Assistant Administrators, Associate Administrators, Regional Administrators, Laboratory Directors, Headquarters Staff Office Directors, General Counsel, and Inspector General shall:

(1) Implement approved data element and data element concept guidelines published under the provision of this policy.

(2) Establish an organization-wide data standards work group that reviews and provides information and comments on proposed data element and data element concept definitions.

7. DEFINITIONS.

a. “Concept” is a unit of thought formed through abstraction of characteristics common to a set of objects.

b. “Data registry” is a database used for data that refers to the use and structure of other data; that is, a database for the storage of metadata.

c. “Data element” is a unit of data for which the definition, identification, representation, and permissible values are specified.

d. “Definition” is a statement that expresses the essential nature of a data element and permits its differentiation from all other data elements.

e. “Object class” is a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behavior follow the same rules.

f. “Property” is a classification of any feature that humans naturally use to distinguish one individual object from another. It is any one of the characteristics of an object class that humans use as a label, quantity, or description.

g. “Qualifier” is a term that helps define and render a concept unique.

8. PROCEDURES. The following appendix to this Order contains information relevant to implementation of the data definition standard. It provides rules and guidelines as well as examples of data element and data element concept definitions.

APPENDIX TO THE SPECIFICATION AND STANDARDIZATION OF DATA ELEMENTS, RULES, AND GUIDELINES FOR THE FORMULATION OF DATA DEFINITIONS

1.0 INTRODUCTION.

1.1 Purpose of Appendix. The purpose of this appendix is to provide further detail on the data element and data element concept definition standard announced in the preceding Order. While the Order introduces the data guideline, it does not contain the full level of detail necessary for programs to form a working understanding of the guideline.

1.2 Background of the Specification and Standardization of Data Elements, Rules, and Guidelines for the Formulation of Data Definitions. The Agency has long striven to create a standard scheme for data definitions that could be used not only in individual programs, but across the Agency as a whole. Such a scheme would help EPA more readily generate responses to public inquiries and determine patterns of compliance behavior across programs.

These reasons, however, are only part of the justification for adopting this new standard. EPA is experiencing a vigorous trend toward data sharing such as site characterizations, risk assessment, and environmental analyses that require the integration of data on individual facilities from diverse sources. The net result is an even greater need for data integration and sharing across different environmental media and programs. This trend is often acknowledged by staff in single-media programs who now face increasing demands for data sharing and integration.

2.0 SCOPE OF DATA DEFINITION STANDARD.

This section augments the discussion of the data element and data element concept definition guideline in an attempt to anticipate questions that program managers may have regarding applicability of the guideline.

2.1 Definition of a data element. Successful implementation of a data element and data element concept definition guideline hinges on a consistent understanding of what “defining a data element” means. The guidelines are specific in the definitions, and a consistent understanding is necessary when deciding which part of the standard to apply to a particular data element and data element concept. An Agency-wide “definition of a data element” is difficult to establish, and, as a result, perceptions of a data element and data element concept definition differ from program to program.

2.2 Rules and Guidelines. The primary characteristics deemed necessary to convey the essential meaning of a particular definition will vary according to the level of generalization or specialization of the data element and data element concept. The primary characteristics should include consideration of the relevance of any object class, property, and qualifiers associated with the concepts being analyzed.

a. The data definition shall:

(1) Be unique (within any data dictionary in which it appears).
(2) Be stated in the singular.
(3) State what the concept is, not only what it is not.
(4) Be stated as a descriptive phrase or sentence(s).
(5) Contain only commonly understood abbreviations.
(6) Be expressed without embedding definitions of other data elements or underlying concepts.

b. A data definition should:

(1) State the essential meaning of the concept.
(2) Be precise and unambiguous.
(3) Be concise.
(4) Be able to stand alone.
(5) Be expressed without embedding rationale, functional usage, domain information, or procedural information.
(6) Avoid circular definitions.
(7) Use the same terminology and consistent logical structure for related definitions.

2.3 Examples of data definitions. To facilitate understanding of the rules for construction of well-formed data element and data element concept definitions, explanations and examples are provided below. Each rule is followed by a short explanation of its meaning. Examples are given to support the explanations. In all cases, a good example is provided to exemplify the explanation. A poor, but commonly used example, is given to show how a definition should not be constructed. The examples are followed by a statement of rationale behind them.

2.3.1 Be Unique - Each definition shall be distinguishable from every other definition (within the dictionary) to ensure that the specificity is retained.

EXAMPLE:
1) Good definitions: “Carbon Dioxide” - A colorless, odorless, non-poisonous gas with the chemical formula $CO2$.

“Carbon Monoxide” - A colorless, odorless, poisonous gas with the chemical formula CO.

2) Poor definitions: “Carbon Dioxide” - A colorless, odorless gas.

. “Carbon Monoxide” - A colorless, odorless gas.

REASON: The poor definition uses the same definition. Each definition must be different.

2.3.2 Be stated in the singular - The concept expressed by the data definition shall be expressed in the singular.

EXAMPLE:
1) Good definition: "Contract Lab" - Laboratory under contract to EPA, which analyzes samples taken from waste, soil, air, and water.

2) Poor definition: "Contract Labs" - Laboratories under contract to EPA, which analyze samples taken from waste, soil, air, and water.

REASON: The poor definition uses the plural word "laboratories."

2.3.3 State what the data element concept is, not only what it is not - The concept cannot be defined exclusively by stating what the concept is not.

EXAMPLE:
1) Good definition: "Sanitary Sewers" - Underground pipes that carry off only domestic or industrial waste.

2) Poor definition: "Sanitary Sewers" - Underground pipes that do not carry storm water.

REASON: The poor definition does not specify what is carried off, only what is not.

2.3.4 Be stated as a descriptive phrase or sentences - A phrase is necessary to form a precise definition that includes the essential characteristics of the concept. Simply stating one or more synonyms or restating the words of the name in a different order is insufficient.

EXAMPLE:
1) Good definition: "Facility" - Something that is built, installed, or established to serve a particular purpose.

2) Poor definition: "Facility" - Building.

REASON: The poor definition contains a near-synonym of the data element name

2.3.5 Contain only commonly understood abbreviations - Understanding the meaning of an abbreviation, including acronyms and initialisms, is usually confined to a certain environment. Exceptions to this rule may be made if the abbreviation is commonly understood such as “i.e.” and “e.g.”

EXAMPLE:

1) Good definition: “Control Technique Guidelines” - A series of EPA documents designed to assist states in defining reasonable available control technology (RACE) for major sources of volatile organic compounds (VOC).

2) Poor definition: “Control Technique Guidelines” - A series of EPA documents designed to assist states in defining RACE for major sources of VOC.

REASON: The good definition contains the meaning of the acronym, therefore the users do not have to refer to other sources to determine what it represents. The acronyms could be left out entirely.

2.3.6 Be expressed without embedding definitions of other data elements or underlying concepts - The definition of another data element or related concept should not appear in the definition of the data element. If the second definition is necessary, it may be an entry in the data registry that is included as a comment.

EXAMPLE:

1) Good definition: “Emissions Trading” - EPA’s policy that allows a plant complex with several facilities to decrease pollution from some facilities while increasing it from others, so long as total results are equal to or better than previous limits.

2) Poor definition: “Emissions Trading” - EPA’s policy that allows a plant complex with several facilities to decrease pollution from some facilities while increasing it from others, so long as total results are equal to or better than previous limits. A facility is something that is built, installed, or established to serve a particular purpose.

REASON: The poor definition contains a definition for facility.

2.3.7 State the essential meaning of the concept - All primary characteristics of the concept represented should appear in the definition at the relevant level of specificity for the context. The inclusion of nonessential characteristics should be avoided. The level of detail necessary is dependent upon the needs of the system user and environment.

EXAMPLE:

1) Good definition: "Tolerances" - The permissible residue levels for pesticides in raw agricultural produce and processed foods.

2) Poor definition: "Tolerances" - The permissible residue levels for pesticides in raw agricultural produce and processed foods. Whenever a pesticide is registered for use on a food or a feed crop, a tolerance (or exemption from the tolerance requirement) must be established. EPA establishes the tolerance levels, which are enforced by the Food and Drug Administration and the Department of Agriculture.

REASON: In the bad definition, nonessential characteristics are included.

2.3.8 Be precise and unambiguous - The exact meaning and interpretation of the defined concept should be apparent from the definition. A definition should be clear enough to allow only one possible interpretation.

EXAMPLE:

1) Good definition: "Coliform Index" - A rating of the purity of water based on a count of fecal bacteria.

2) Poor definition: "Coliform Index" - A rating of the purity of water based on fecal bacteria.

REASON: The poor definition does not illustrate what determines the Coliform Index "rating."

2.3.9 Be concise - The definition should be brief and comprehensive.

EXAMPLE:

1) Good definition: "Aeration Tank" - A chamber used to inject air into water.

2) Poor definition: "Aeration Tank" - A chamber used to inject air into water which promotes the degradation of organic water. The process may be passive or active.

REASON: In the poor definition, all the phrases after "…which promotes " are extraneous qualifying phrases.

2.3.10 Be able to stand alone - The meaning of the concept should be apparent from the definition.

EXAMPLE:

1) Good definition: “Sewage Lagoon” - A shallow pond where sunlight, bacterial action, and oxygen work to purify wastewater.

2) Poor definition: “Sewage Lagoon” - A lagoon used to purify wastewater. See “lagoon.”

REASON: The good definition does not require the aid of a second definition (lagoon) to understand the meaning of the first.

2.3.11 Be expressed without embedding rationale, functional usage, domain information, or procedural information - Although they are often necessary, such statements do not belong in the definition proper because they contain information extraneous to the purpose of the definition.

EXAMPLE:

1) Good definition: “Best Available Control Technology” - An emission limitation based on the maximum degree of emission reduction.

2) Poor definition: “Best Available Control Technology” - An emission limitation based on the maximum degree of emission reduction which (considering energy, environmental, and economic impacts and other costs) is achievable through application of production processes and available methods, systems, and techniques.

REASON: The good definition does not contain remarks about functional usage, e.g., “which (considering energy...).”

2.3.12 Avoid circular definitions - Two definitions shall not be defined in terms of each other. A definition should not use another concept’s definition as its definition. The following example illustrates this.

EXAMPLE:

1) Good definition: "Facility Identifier" - A unique number assigned to something that is built, installed, or established to serve a particular purpose.

2) Poor definitions: “Facility Identifier” - A unique number assigned to a facility.

“Facility” - Building corresponding to the facility id number.

REASON: The poor definition refers to the other for its meaning. The meaning is not given in either definition.

2.3.13 Use the same terminology and consistent logical structure for related definitions - A common terminology and syntax should be used for similar or associated definitions.

EXAMPLE:

1) Good Definitions:

"LD 0" - The highest concentration of a toxic substance at which none of the test organisms die.

"LD L0" - The lowest concentration of a toxic substance which kills all test organisms.

2) Poor Definition:

"LD L0" - All test organisms are killed when the concentration of a toxic substance is at the lowest possible level.

REASON: The good definition uses the same terminology and the syntax facilitates understanding. Otherwise, users wonder whether some difference is implied by use of variable syntax.